doi:10.1016/j.jmb.2005.09.085 J. Mol. Biol. (2005) 354, 994–1007

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

SVA Elements: A Hominid-specific Retroposon Family

Hui Wang†, Jinchuan Xing†, Deepak Grover†, Dale J. Hedges Kyudong Han, Jerilyn A. Walker and Mark A. Batzer*

Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-Scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge LA 70803, USA

**Corresponding author*

SVA is a composite repetitive element named after its main components, SINE, VNTR and *Alu*. We have identified 2762 SVA elements from the human genome draft sequence. Genomic distribution analysis indicates that the SVA elements are enriched in G+C-rich regions but have no preferences for inter- or intragenic regions. A phylogenetic analysis of the elements resulted in the recovery of six subfamilies that were named SVA_A to SVA_F. The composition, age and genomic distribution of the subfamilies have been examined. Subfamily age estimates based upon nucleotide divergence indicate that the expansion of four SVA subfamilies (SVA_A, SVA_B, SVA_C and SVA_D) began before the divergence of human, chimpanzee and gorilla, while subfamilies SVA_E and SVA_F are restricted to the human lineage. A survey of human genomic diversity associated with SVA_E and SVA_F subfamily members showed insertion polymorphism frequencies of 37.5% and 27.6%, respectively. In addition, we examined the amplification dynamics of SVA elements throughout the primate order and traced their origin back to the beginnings of hominid primate evolution, approximately 18 to 25 million years ago. This makes SVA elements the youngest family of retroposons in the primate order.

Keywords: SVA; subfamily; genomic distribution; polymorphism; phylogenetic distribution

Introduction

Transposons and transposon-like repetitive elements collectively occupy about 44% of the human genome. *Alu* and L1 (long interspersed element-1) elements account for ~30% of the genome sequence and are the most abundant transposable elements in humans,[1] while human endogenous retroviruses (HERVs) represent another ~1% of the human genome. In addition to the major retrotransposon families, there are smaller families of transposons such as SVA, which are receiving increased attention lately due to their recent retrotransposition activity in the human genome.[2,3]

The SVA element was originally named SINE-R, with the R indicating its retroviral origin.[4] In 1994, Shen *et al.* identified a new composite retroposon when they studied the structure of the RP gene.[5] This new retroposon consisted of the SINE-R element together with a stretch of sequence that shares sequence similarity with *Alu* sequences. Thus, it was named "SVA" after its main components, SINE, VNTR and *Alu*.[5]

SVA elements contain the hallmarks of retrotransposons, in that they are flanked by target site duplications (TSDs), terminate in a poly(A) tail and they are occasionally truncated and inverted during their integration into the genome.[2] In addition, they can transduce 3′ sequences during their movement. Therefore, it has been proposed that SVA elements are non-autonomous retrotransposons that are mobilized by L1 encoded proteins *in trans*.[2]

SVA elements remain active in the human genome, as demonstrated by their involvement in the creation of various diseases. To date, at least four diseases have been reported related to SVA insertions.[2,6–10] This makes SVA the third known category of retrotransposons currently expanding in the human lineage, along with L1 and *Alu* elements.[11,12]

To assess the distribution and impact of SVA elements in the human genome, we examined all

† H.W., J.X. and D.G. contributed equally to this work.

Abbreviations used: VNTR, variable number of tandem repeat; HERV, human endogenous retrovirus; TSD, target site duplication; SINE, short interspersed element; LTR, long terminal repeat; QPCR, quantitative PCR; pol, RNA polymerase.

E-mail address of the corresponding author: mbatzer@lsu.edu

the SVA elements in the human genome reference sequence.[1] Six SVA subfamilies were identified and characterized. For the two human-specific subfamilies, the associated insertion presence/absence human genomic diversity was analyzed. Furthermore, we traced the origin of the entire SVA family back to the beginning of the hominid primate radiation and determined the copy number of SVA elements in different non-human primate genomes. The overall distribution of SVA elements showed a significant correlation with genomic G+C content and gene density.

Results

Copy number and genomic distribution

Copy number of SVA elements

In total, 2762 SVA elements were identified in the human genome draft sequence (hg17; May 2004 freeze). Together, they account for 4.2 Mb of the genome, with an average density of one element per 1.03 Mb. Among them, 1752 elements are full-length, composing 63% of the group. The copy number of SVA elements in the chimpanzee genome (panTro1; Nov 2003 freeze) was also determined. A total of 2637 elements were obtained, with 42% being full-length. After examining the truncated SVA elements in the chimpanzee genome draft sequence, we found that a large proportion of the SVA elements were truncated by stretches of Ns, corresponding to unsequenced/unassembled regions of the genome. Therefore, the lower proportion of full-length elements in the chimpanzee genome may be due to the lower quality of the chimpanzee draft sequence compared to its finished human counterpart. Detailed wet bench estimates of the copy number and phylogenetic distribution of SVA elements throughout the primate order are outlined below.

The structure of the SVA element

The full-length SVA element can be divided into five components (Figure 1(a)): (1) a $(CCCTCT)_n$ hexamer simple repeat region which is located at the 5′ end; (2) an *Alu* homologous region, usually composed of two antisense *Alu* fragments and an additional sequence of unknown origin; (3) a variable number of tandem repeat (VNTR) region, composed of a variable number of copies of a 35–50 bp sequence; (4) a short interspersed element (SINE) region about 490 bp long, which is derived from the 3′end of the *env* gene and the 3′ long terminal repeat (LTR) of the endogenous retrovirus HERV-K10;[4] and (5) a poly(A) tail after a putative polyadenylation signal (AATAAA). We extracted the VNTR region from all of the full-length SVA elements and analyzed their length variation (Figure 1(b)). The length of the SVA VNTR region varies from 48 bp to 2306 bp, with an average value of 819 bp. Two-thirds of the SVA elements have VNTR lengths in the range of 400–900 bp.

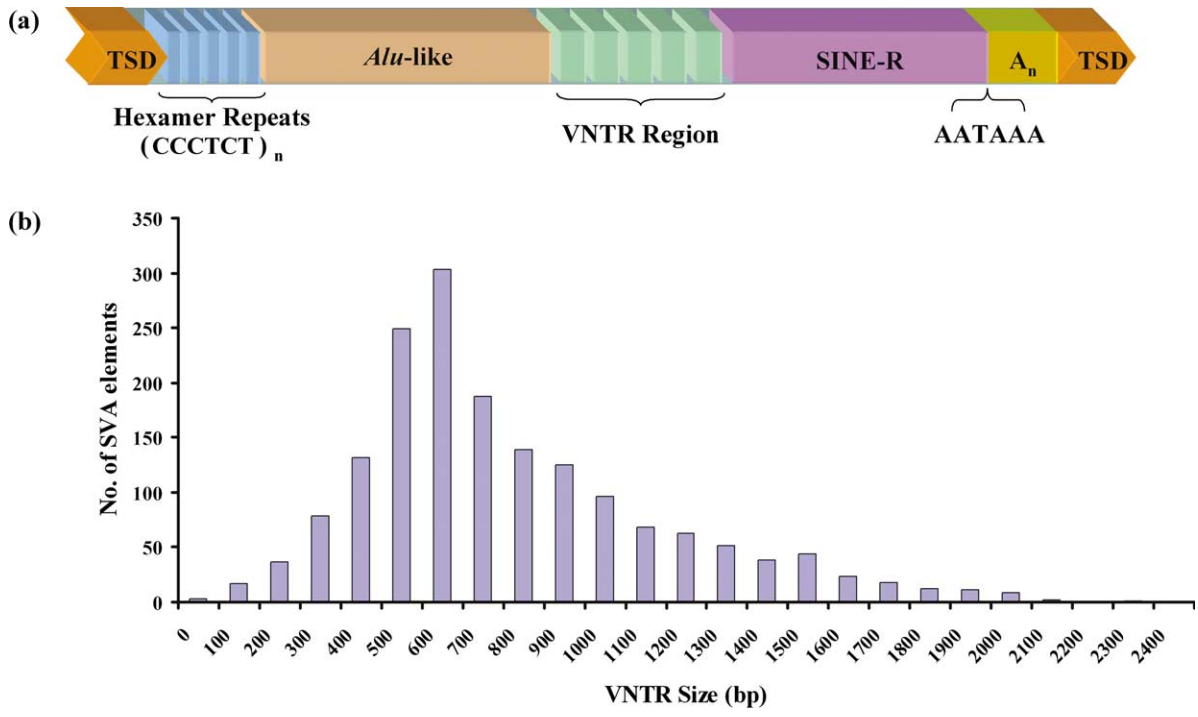

Figure 1. (a) Structure of a full-length SVA element with target site duplications. Various regions of the SVA element are color-coded and denoted. (b) Size distribution of SVA VNTR regions. The VNTR regions of SVA elements are shown in 100 bp intervals or bins.

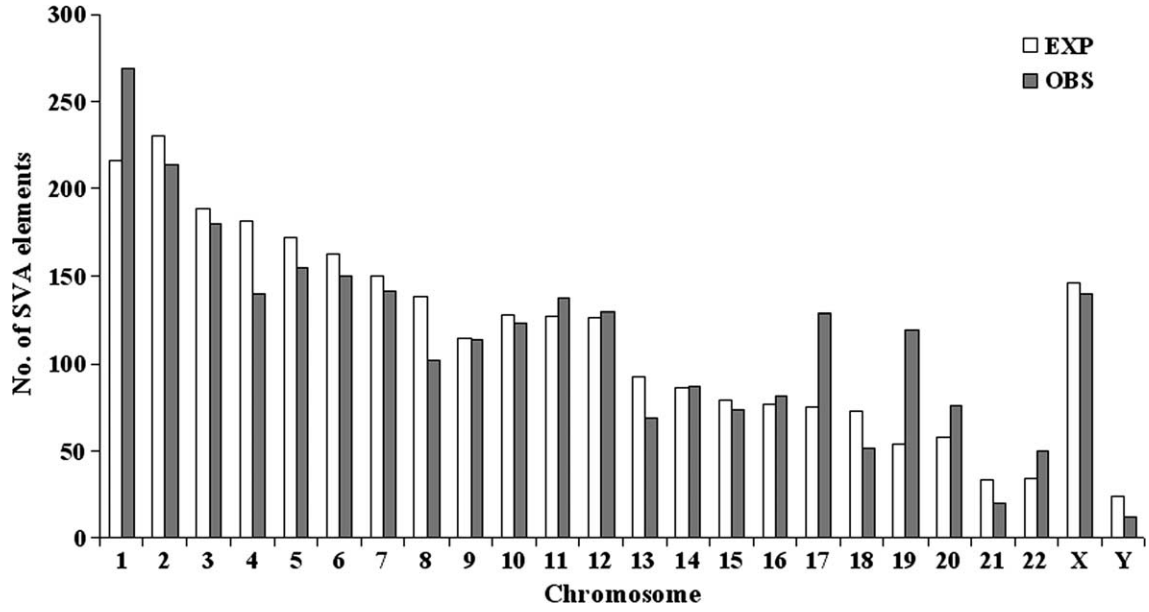

Figure 2. Human genomic distribution of SVA elements. The observed and expected number of SVA elements from each chromosome are shown. The expected numbers on each chromosome were obtained by multiplying the chromosomal length with average density of these elements in the genome.

Genomic distribution

To examine the genomic distribution of SVA elements, we first analyzed their distribution at the chromosomal level by comparing the observed number of the elements with the expected numbers, assuming an infinite sites (random) insertion model (Figure 2). In this model, the number of insertions on each chromosome was solely proportional to the size of the chromosome. A chi-square analysis revealed that the two distributions were significantly different ($\chi^2 = 78.29$, $df = 23$, $p < 0.001$), thus leading us to reject the simple random insertion model. In particular, the density of SVA elements was found to be much higher than expected on chromosomes 1, 17, 19 and 22. On the other hand, chromosomes 4, 5, 13, 18, 21 and Y had far fewer elements than expected. A similar non-random chromosomal distribution has previously been observed for *Alu* elements.[13,14]

Since the properties of the genomic sequence vary greatly among different human chromosomes,[1] we further analyzed the distribution of SVA elements in relation to other genomic properties, such as G+C content, gene content and repeat content of the flanking nucleotide sequence. At the whole-genome level, we observed a positive correlation between SVA density and both G+C content ($r = 0.59$; $p = 0.002$) and gene density ($r = 0.53$; $p = 0.007$). In terms of distribution, the SVA elements resemble *Alu* elements, which are also preferentially found in high G+C/gene-rich regions of primate genomes, although the extent of correlation is much higher in the case of *Alu* elements.[1] In terms of vicinity to genes, we observed that 1025 SVA elements reside in intronic regions of the genes and 1737 elements occur in intergenic regions. Statistically, there is no significant difference ($\chi^2 = 11.48$, $df = 23$, $p = 0.98$) between the number of SVA elements in these two regions. This result suggests that there is no preference for SVA insertion in genes or intergenic regions.

To examine the SVA distribution in relation to the global repeat content, we extracted nucleotide sequence intervals of 1, 2, 5, 10, 25 and 50 kb from both 5′ and 3′ flanking regions of SVA elements and analyzed the repeat content using RepeatMasker† (Table 1). We found that the repeat content in these different-sized flanking sequences were not significantly different from each other or the overall repeat content in the genome ($p > 0.2$ in each case). Then, we studied individually the two most abundant repeat families, *Alu* and L1 in these flanking sequences. When compared with their expected distribution frequency throughout the genome, *Alu* elements were found to be overly represented in these regions, whereas L1 elements were underrepresented (in both cases, $p < 0.00001$). A closer analysis of the SVA flanking regions revealed that the G+C content in the flanking regions is much higher than the genomic average ($p < 0.00001$). Given high *Alu* density in G+C-rich, and high L1 density in A+T-rich regions of the genome,[1] these results are not surprising and indicate that SVA elements are not preferentially distributed in repeat rich regions of the genome.

Subfamily analysis

Subfamily structure and composition

If SVA elements expanded in a process similar to *Alu* and L1 elements, a hierarchical subfamily

† http://www.repeatmasker.org

Table 1. Repeat content and G+C content in the SVA flanking regions

	Total repeat (%)	*Alu* (%)	L1 (%)	G+C (%)
1 kb	43.50	12.40	12.25	42.05
2 kb	43.97	12.79	12.62	41.91
5 kb	44.18	12.79	12.76	42.85
10 kb	44.08	12.82	12.74	42.84
25 kb	43.91	13.02	12.67	43.00
50 kb	45.12	13.00	12.52	43.06
Whole genome	44.83	10.60	16.89	40.91

structure should be formed. To identify possible subfamily structure, multiple alignments of the SVA elements were constructed. Due to the highly variable VNTR region and considerable number of 5′ truncated elements, only the LTR derived region (referred to as the S part in the following text) was used for the initial analysis. Examination of the alignments resulted in the recovery of at least six different groups among the elements based on their diagnostic substitutions. Consensus sequences for these groups of SVA elements were generated in BioEdit v7.0[15] using a "majority rules" approach. Some of the CpG sites were reconstructed manually due to their high mutation rate.[16,17] These subfamilies were named SVA_A to SVA_F. Using a similar approach, one chimpanzee-specific subfamily was identified and named SVA_PtA. The lineage specific distribution of this SVA subfamily was verified by BLAT (blast-like alignment tool)[18] comparison to the human genome.

To validate the subfamily definition, multiple alignments were also constructed using the *Alu* related region (referred to as the A part in the following text) together with the S part and the same groups were identified. Thus, we constructed the human SVA subfamily consensuses including both the A and S regions of the element. The multiple alignments of the human SVA subfamily consensuses along with the corresponding HERV-K10 sequences are available in the supplemental alignment†.

Next, the relative proportions of each SVA subfamily were calculated (Figure 3). Among all the subfamilies, SVA_D represents the largest subfamily and accounts for over 40% of the family. The second largest subfamily, SVA_B, accounts for about 15% of the family. The remaining subfamilies (A, C, E and F) have relatively small numbers of elements (<300). Since we recognized the SVA subfamilies based on the intact S part of the elements, we could not group the elements into subfamilies when they lacked a full-length S part. These truncated elements were collectively designated "others" and accounted for 15% of the family.

SVA subfamily age estimates

The ages of different SVA subfamilies were estimated using a method similar to that used for

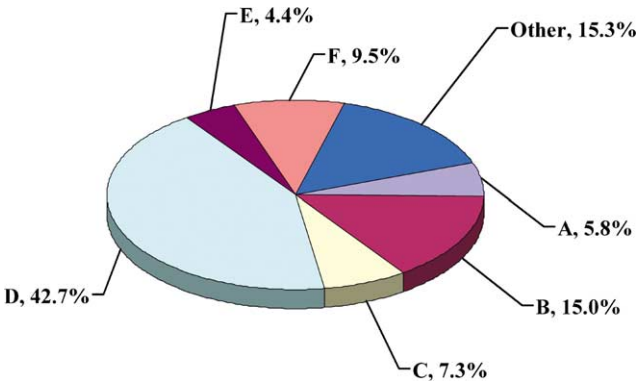

Figure 3. SVA subfamily composition in the human genome. The breakdown of SVA elements into various subfamilies is shown as a percentage of all the SVA elements in the human genome.

Alu subfamilies described previously.[16] Briefly, the S parts of the SVA elements in each subfamily were aligned with the subfamily consensus. Next, substitutions in this region were divided into CpG and non-CpG substitutions and substitution density was calculated using a Perl script. Separate neutral substitution rates of 0.0015/site per million years (Myrs) and 0.0090/site per Myrs were used for non-CpG and CpG substitutions, respectively.[16] The substitution density and age estimate of each subfamily are shown in Table 2. Given the approximate divergence time of hominid primates,[19,20] the age estimates indicate that subfamily SVA_A (13.56 Myrs) may have expanded contemporary to the divergence of the orangutan and the great apes (human, chimpanzee and gorilla) (12–15 million years ago (Mya)). The expansion of subfamilies SVA_B (11.56 Myrs), SVA_C (10.88 Myrs) and SVA_D (9.55 Myrs) may have predated the human, chimpanzee and gorilla divergence (~7 Mya). The relatively young age of subfamilies SVA_E (3.46 Myrs) and SVA_F (3.18 Myrs) suggested these two subfamilies may have expanded after the human and chimpanzee divergence (~4–6 Mya). Indeed, when the human and chimpanzee sequences were compared using BLAT, the members of subfamilies SVA_E and SVA_F were absent at chimpanzee orthologous loci, confirming their human-specific distribution.

Having obtained age estimates for the SVA subfamilies, we examined the properties of SVA VNTR repeat regions as a function of subfamily and age. ANOVA results indicate that there are significant ($p \ll 0.001$) differences among subfamilies for VNTR lengths. With respect to age, if the oldest SVA subfamily (SVA_A) is excluded from the analysis, we find a significant negative correlation ($r^2 = 0.96$) of VNTR lengths with time‡. It is unclear why this

† Available at http://batzerlab.lsu.edu, under publications.

‡ Supplemental Figure 1 available at http://batzerlab.lsu.edu, under publications.

Table 2. Age estimates of SVA subfamilies

SVA subfamily	CpG density (%)	Non-CpG density (%)	Age CpG (Myrs)	Age non-CpG (Myrs)	Age average (Myrs)
SVA_A	15.13	1.55	16.81	10.30	13.56
SVA_B	9.53	1.88	10.59	12.53	11.56
SVA_C	9.73	1.64	10.81	10.94	10.88
SVA_D	8.50	1.45	9.45	9.64	9.55
SVA_E	2.21	0.67	2.46	4.47	3.46
SVA_F	2.46	0.54	2.73	3.63	3.18

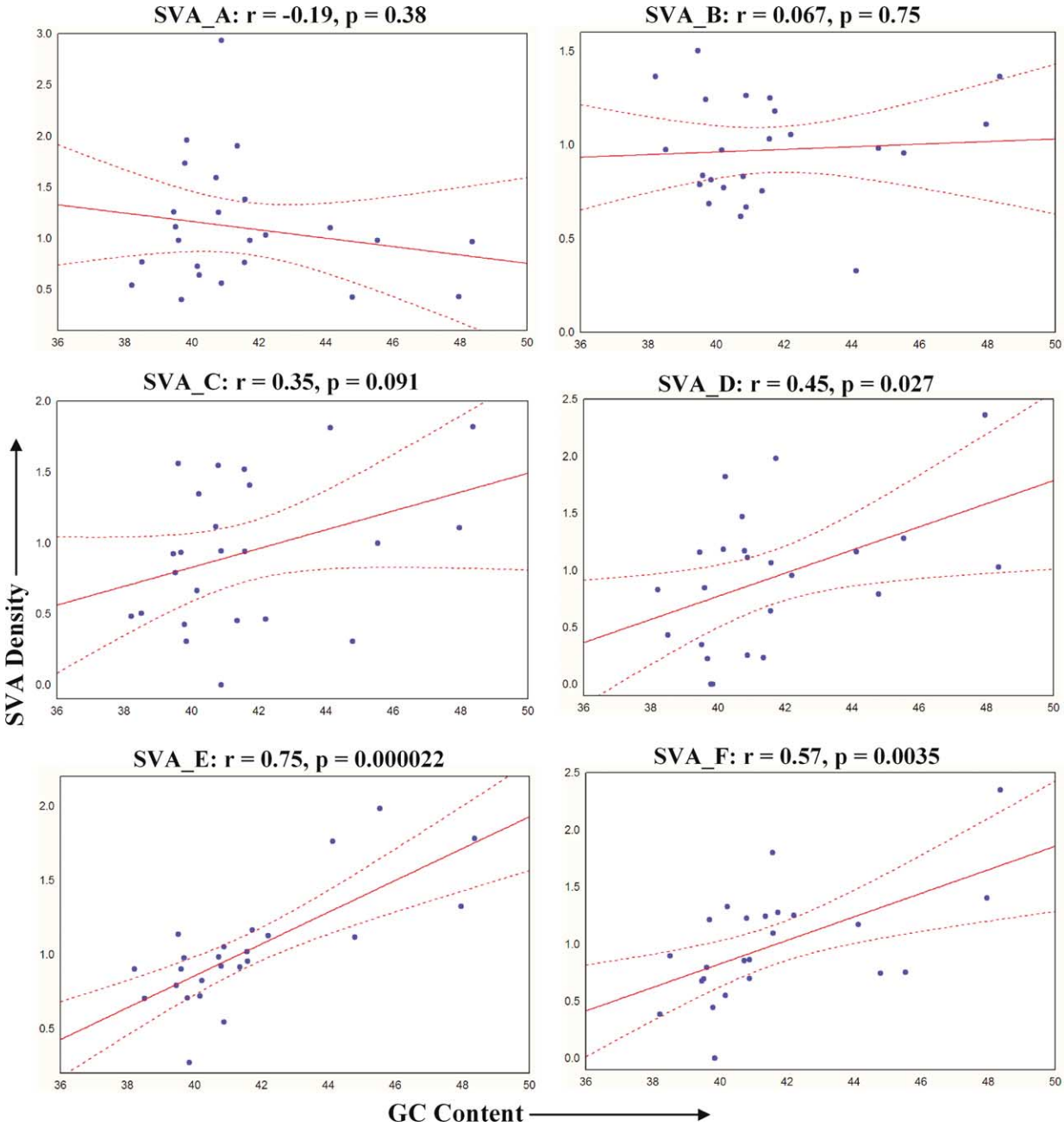

Figure 4. Correlation between SVA subfamily distribution and chromosomal G+C content. The linear regression is denoted with a continuous line and the 95% confidence intervals are denoted by the broken lines. Correlation coefficients (*r*) and *p* values are shown.

subfamily deviates from the overall pattern observed.

G+C content distribution, gene density and repeat density in the flanking regions of the subfamilies

A correlation analysis between SVA content and G+C content of all human chromosomes was performed to examine the distributions of different SVA subfamilies (Figure 4). Different levels of correlation were observed among subfamilies, with the highest values for the youngest subfamilies SVA_E ($r=0.75$; $p=0.000022$) and SVA_F ($r=0.57$; $p=0.0035$), mild correlation for subfamily SVA_D ($r=0.45$; $p=0.027$), and no significant correlation for SVA_A, SVA_B and SVA_C. The general trend suggests the enrichment of the youngest SVA subfamilies on the high G+C content chromosomes. This distribution is in contrast to *Alu* and L1 elements, whose younger subfamilies are preferentially found in A+T-rich regions.[1,21]

To examine whether the distribution of SVA subfamilies observed at the chromosomal level also exists in different regions within a chromosome, the human genome was separated into 12 bins based upon their G+C content[21] and the SVA density in each of the bins was calculated. We pooled the subfamilies into three groups according to their ages: old subfamilies (A, B and C), the intermediate subfamily (D) and young subfamilies (E and F). Their distributions in relation to the genomic G+C content were plotted (Figure 5). As shown in the figure, SVA elements reside mainly in medium to medium-high G+C-rich regions of the genome, with maximum density in the bins that correspond to 42–50% G+C content. However, there was a shift in SVA density towards higher G+C bins with a decrease in evolutionary age. The older subfamilies (A, B and C) were quite rich in the 40–42% G+C bin, where the members of young subfamilies (E and F) were found less frequently. The opposite was observed in 48–52% G+C content regions of the genome, which were extremely rich in young SVA elements. The SVA_D subfamily exhibited an intermediate pattern for density in these G+C bins.

The correlation analysis between SVA subfamilies and chromosomal gene content were examined and similar results were obtained as in the case of G+C content. Strong correlations were obtained from young subfamilies SVA_E ($r=0.77$; $p<0.0001$) and SVA_F ($r=0.64$; $p<0.001$); moderate correlation existed for subfamily SVA_D ($r=0.54$; $p=0.007$) and SVA_C ($r=0.47$; $p=0.02$). The correlations were very low and insignificant for the oldest members of this repeat family (SVA_A and SVA_B) (Figure 6). This result is not surprising, given that gene densities and G+C content are highly correlated in the human genome.[1]

We next analyzed the repeat distribution in the flanking regions of SVA elements belonging to each subfamily (Figure 7). The variations in *Alu* and L1 contents across different subfamilies correspond with the G+C content of the flanking regions. In fact, we observed that *Alu* and L1 element density in these regions is significantly correlated with the G+C and A+T content, respectively ($p>0.001$). These data suggest that SVA elements are not preferentially distributed in either *Alu*-rich or L1-poor regions of the genome.

Human genomic diversity of two human-specific subfamilies

To assess the human genetic diversity related to the SVA elements, elements from SVA_E and SVA_F subfamilies were screened for insertion polymorphism on a diverse human population panel

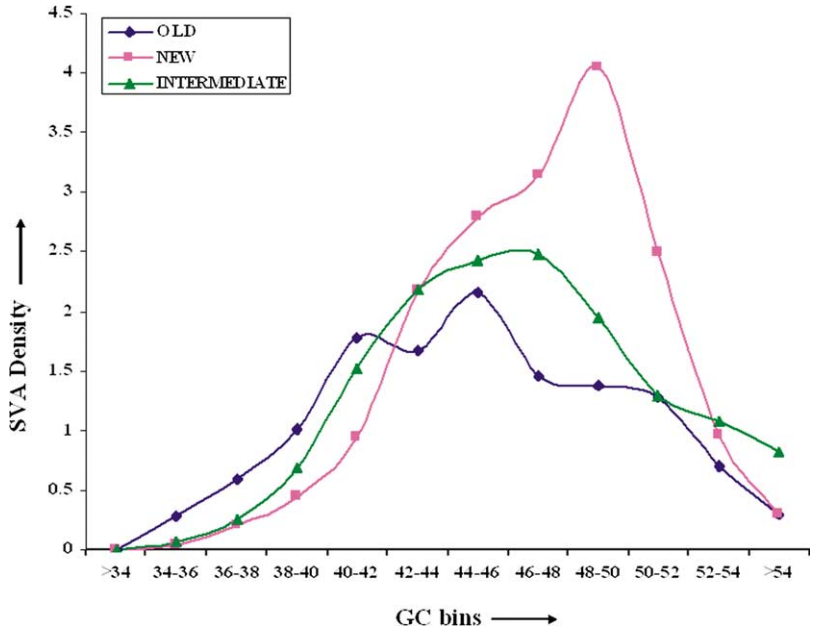

Figure 5. SVA subfamily and genomic G+C content. The bin separations are identical with those previously described by Lander *et al.* [1]SVA subfamilies are pooled into old (SVA_A, SVA_B, SVA_C), intermediate (SVA_D) and young (SVA_E and SVA_F) subfamilies.

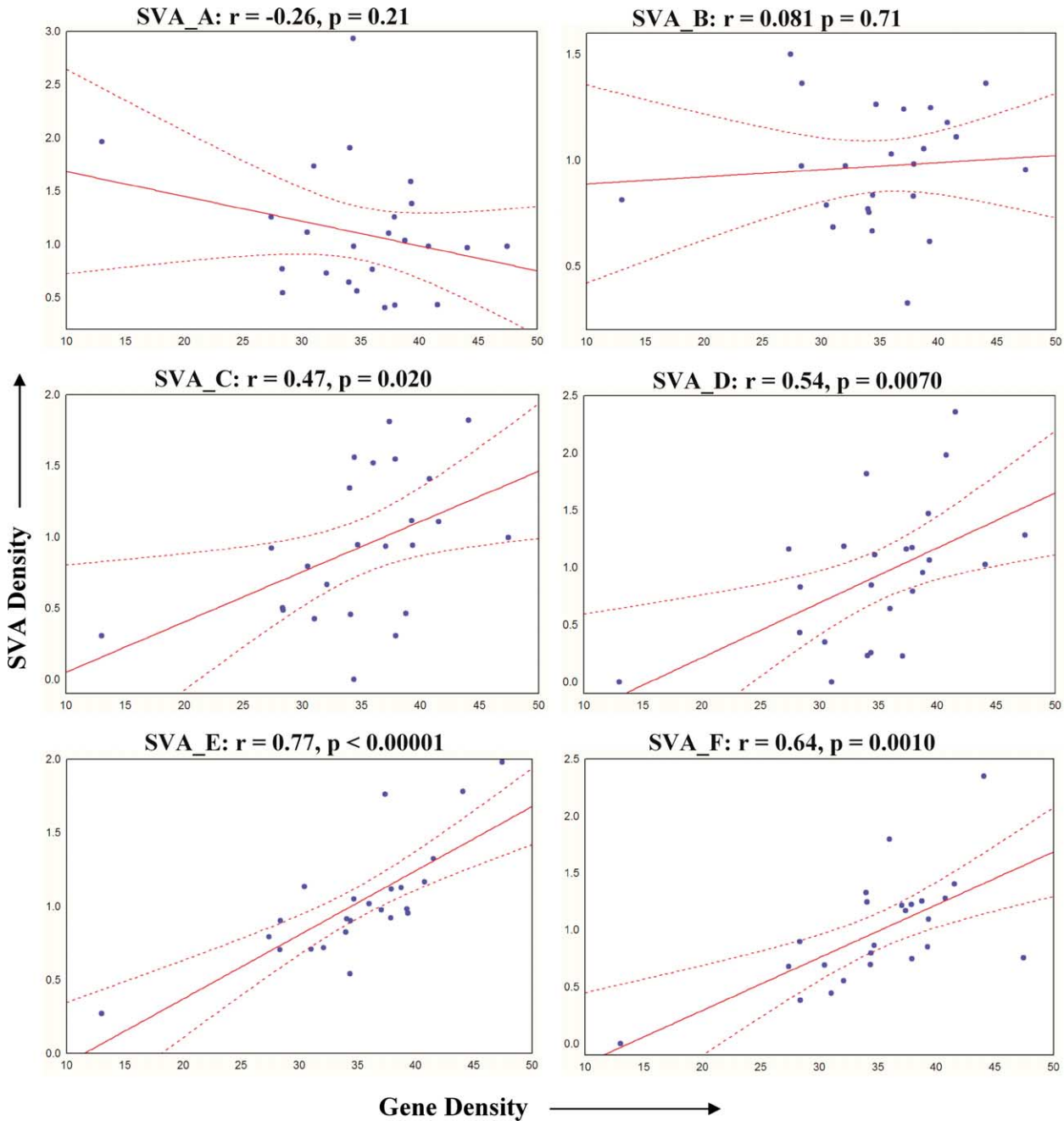

Figure 6. Correlation between SVA subfamilies and chromosomal gene density. The linear regression is denoted with a continuous line and the 95% confidence intervals are denoted by the broken lines. Correlation coefficients (r) and p values are shown.

(see Materials and Methods). In total, 48 members of subfamily SVA_E and 58 members of SVA_F were surveyed. For SVA_E, 37.5% (18/48) of the elements showed insertion presence/absence polymorphism on our panel with a 27.6% (16/58) polymorphism rate for SVA_F. The insertion polymorphism rates of these two subfamilies are comparable to the human-specific *Alu* and L1 subfamilies with similar ages,[14,22–24] further demonstrating the contemporary retrotransposition activity of the SVA family of retroposons. The detailed insertion allele frequencies, heterozygosities and genotypes for the SVA insertion polymorphisms are shown in supplemental Table 1.†

Evolution of SVA elements

Evolution of SVA subfamilies

To examine the relationship among the subfamilies, a median-joining network[25,26] was constructed

† Available at http://batzerlab.lsu.edu, under the publications section.

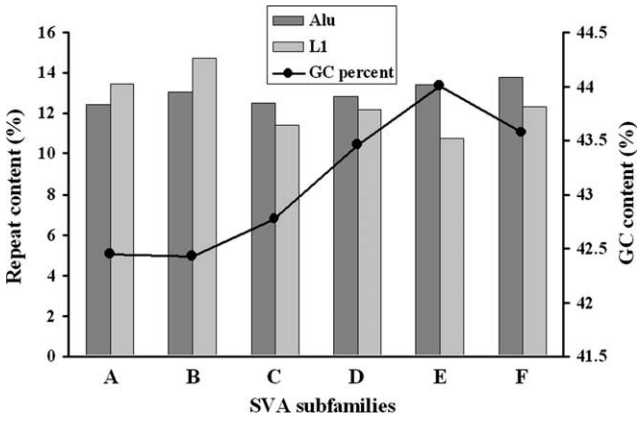

Figure 7. L1 and *Alu* densities flanking SVA elements and G+C content. The percentage of *Alu* and L1 elements flanking SVA elements from different subfamilies are shown by the dark bars. The flanking unique sequence G+C content is shown by the line among SVA subfamilies.

using the S part of the subfamily consensuses and the corresponding region of HERV-K10 (Figure 8). The network analysis indicates that the older SVA subfamilies evolved in a single lineage: the SVA_A consensus has the highest sequence similarity to the HERV-K10 counterpart, differing by 27 substitutions. SVA_B differs from SVA_A by nine substitutions and a 16 bp deletion is present at the 5′ end of the SVA_B consensus as compared to HERV-K10 sequence and the SVA_A consensus. This deletion is present in all other subfamily consensuses as well. SVA_C is derived from SVA_B and differs from SVA_B by ten substitutions, while SVA_D differs from SVA_C by 15 substitutions. Unlike the older subfamilies, the human-specific SVA_E and SVA_F derived independently from a re-constructed ancestral sequence, which is only one substitution different from SVA_D and they differ from SVA_D by 15 and 14 substitutions, respectively.

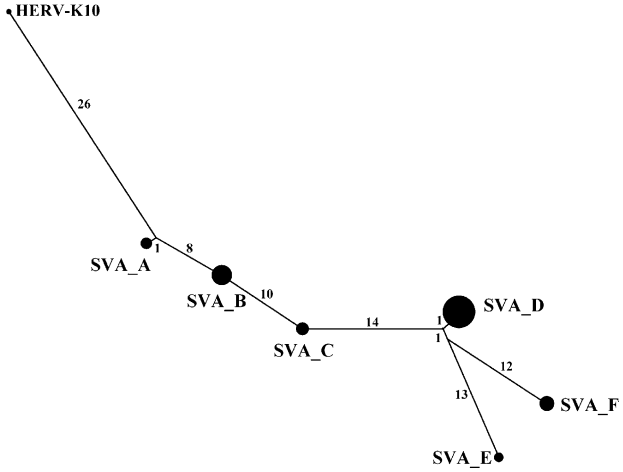

Figure 8. Median-joining network of the SVA subfamilies. The network of the SVA subfamily was reconstructed using the S part of the consensus sequences from each subfamily along with the corresponding region of HERV-K10. The lines denote substitution steps, with the number of substitutions shown on the top of the lines. The size of each of the circles corresponds to the relative size of the subfamily in the human genome.

Copy number of SVA elements in different non-human primate genomes

To further investigate the evolutionary history of the SVA elements, the copy number of SVA elements in different primate genomes was estimated by quantitative PCR (QPCR) using a pair of intra-SVA primers. QPCR results from different primates were normalized based on the SVA copy number in the human genome (2762) (Figure 9). The result indicated that both common chimpanzee (*Pan troglodytes*) (2769) and pygmy chimpanzee (*Pan paniscus*) (2646) had a similar number of elements compared to human, while gorilla (*Gorilla gorilla*) (2334) appeared to have about 400 fewer SVA elements. Nevertheless, given the standard deviations in our copy number estimates (90, 163 and 221 in common chimpanzee, pygmy chimpanzee and gorilla, respectively), the number of SVA elements in human, chimpanzees and gorilla may not be appreciably different. By contrast, orangutan (*Pongo pygmaeus*), which diverged from gorilla about 12–15 Myrs ago,[19,20] has fewer than 1000 SVA elements. Siamang (*Hylobates syndactylus*) has about 40 elements and no SVA elements were detected in any of the two Old World monkeys (green monkey (*Chlorocebus aethiops*), rhesus macaque (*Macaca mulatta*)) we examined.

To verify the QPCR results, the available Old World monkey genomic sequences in the NCBI database were searched for SVA elements using

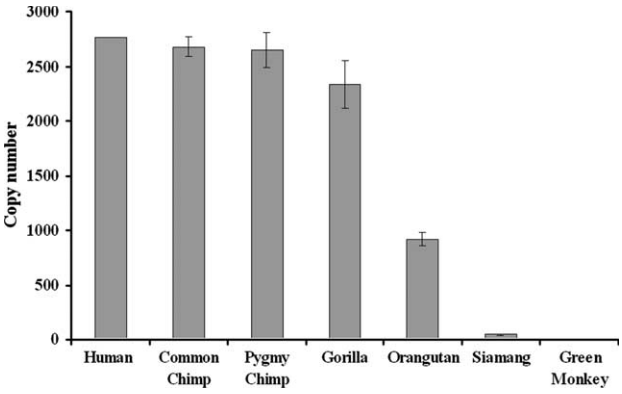

Figure 9. SVA copy number in the primate lineage. The signal intensity derived from the number of known SVA elements in the human genome (2762) was used as standard for the estimations from quantitative PCR experiments. These experiments utilized QPCR of the S part of each element to estimate the copy number as outlined in the Materials and Methods. The standard deviation (SD) of three replicates for each estimation is shown in the Figure as a bar.

BLAST (basic local alignment search tool).[27] In addition, the rhesus macaque draft assembly, Mmul_0.1 (UCSC version rheMac1) was searched using BLAT. In agreement with the QPCR results, no authentic SVA element was identified from either of the database searches. It is noteworthy that the individual components of the SVA elements (*Alu* region, VNTR region and LTR-derived region) are present in the Old World monkey genomes, although the origin or the "composition" of the full-length SVA element appears to have happened after the divergence of hominid and Old World primates.

Furthermore, a PCR display assay was performed as described[28] on the two Old World monkeys (rhesus macaque and green monkey) and four hominid primates (pygmy chimpanzee, common chimpanzee, gorilla and gibbon (*Hylobates lar*)), using primers specific for the A and S parts of the SVA element (SVA_*Alu*: 5′-ATTGAGCACTGAGTGAACGAGA−3′ and SVA_SINE: 5′- AGTACCCAGGGACACAAACACT-3′). Although SVA elements were readily isolated from all of the hominid primates, none was recovered from either of the two Old World primates.

Discussion

Copy number and subfamily composition of SVA elements

The total number of SVA elements reported here was based on a whole-genome analysis; therefore it is more likely to represent the actual number of SVA elements in the human genome compared to previous studies.[2,4] It should be noted here that due to the polymorphic nature of some of the young SVA elements, a proportion of the total number of elements will not be retrieved from any single reference genome that has been sequenced.[29] Therefore, the copy number reported here represents the lower limit of the number of SVA elements in the human genome. For the common chimpanzee genome, a similar number of elements were identified. However, a large proportion of the SVA elements were truncated or flanked by Ns, preventing further investigation of the SVA family in the chimpanzee genome. With improvement in the assembly of the common chimpanzee genome, a more thorough examination of the SVA elements in chimpanzees can be achieved in the future.

Multiple alignment of the S part of the SVA elements revealed at least six subfamilies based on their diagnostic mutations. Aside from the SVA_B and SVA_D subfamilies, the rest of the subfamilies have relatively small numbers of elements (<300). The smaller size of a particular subfamily could be the result of a relatively short amplification period or the lower retrotransposition activity of the source gene(s). The complex nature of the amplification of retrotransposons has recently been underscored by the "stealth model" of *Alu* amplification.[28] It remains a possibility that SVA subfamilies with currently low copy numbers are simply currently in a state of retrotranspositional quiescence.

Truncated SVA elements

SVA elements were initially named SINE-R due to their close relationship to the endogenous retrovirus HERV-K10.[4] The question remained whether the S part of the element first originated from HERV-K10 and expanded alone similar to LINE-mediated integration of HERV-W[30] and the recruitment of the A and VNTR parts happened later, or the SVA element started the expansion as a whole. To address this issue, we examined all of the truncated SVA elements. By comparing the sequence of truncated elements with the subfamily consensus, we found truncated elements distributed among all the subfamilies we identified. This suggests that SVA elements expanded as one composite unit. Furthermore, we examined the alignments of the two oldest subfamilies, SVA_A and SVA_B. The results showed that the majority (>70%) of the elements were full-length, containing the A part of the elements. These two lines of evidence suggest that the SVA elements were fully assembled/composed very early during their expansion as a repeated DNA sequence family. One very interesting possibility is that the acquisition of the antisense *Alu* fragments and the hexamer simple repeats changed the properties of the element (e.g. the hexamer simple repeats provided a promoter for the element) and initiated the expansion of the SVA family. Additional evidence, including cell culture based assays need to be gathered to test this hypothesis.

The mobilization of SVA elements

Sequence examination of SVA elements showed that, unlike *Alu* elements, SVA elements do not have an RNA polymerase III (pol III) internal promoter. In addition, the full-length transcript (~2 kb) is too long for normal RNA pol III transcription. As such, it seems likely that the SVA element is transcribed by RNA polymerase II (pol II), similar to the L1 elements. Indeed, several lines of evidence support this concept: first, SVA has a putative polyadenylation site and the poly(A) tail is added to the element during the retrotransposition. Two types of integrations clearly showed the addition of poly(A) tails: (1) in elements with 3′ transduced sequence, the poly(A) tail at the end of the transduced sequence was not in the genomic sequence but was added post-transcriptionally; (2) in 3′ truncated elements, an alternative polyadenylation site in the SVA element sequence was used so the transcripts of these elements were truncated and the poly(A) tail was added after the alternative site. Second, an extra G residue was present at the beginning of about one-third of the SVA elements. This extra G may represent the 5′ capping

modification of the SVA RNA sequence by RNA pol II. Similar addition of G residues has been observed for L1 elements[31] and reverse transcriptase has been known to reverse transcribe the 5′ cap structure.[32,33]

If SVA elements are indeed transcribed by RNA pol II, do they have their own promoters or do they rely on the promoter activity of their flanking regions? We believe SVA elements may rely on both approaches. At least 10% of SVA elements have 3′ transductions and can be traced to different origination points. The possibility that most of these several hundred elements have fortuitously integrated right after a viable promoter is remote. We believe that these active elements bear a functional promoter region and are capable of transcription by themselves. By contrast, there is at least one 5′ transduced SVA element, which suggests that SVA may be able to use a 5′- flanking promoter.

Genomic distribution of SVA elements

Initial analysis suggested that the genomic distribution of SVA elements was more similar to *Alu* elements than to L1 elements. However, when we studied each SVA subfamily individually, apparent differences were observed between the SVA and *Alu* element distributions. The distribution of the youngest SVA elements (<5 Myrs) showed an apparent enrichment in the G+C-rich regions in the genome, with a peak at 48–50% G+C content. When the age of the SVA elements increased, the enrichment of the elements became less apparent and the peak of distribution shifted towards the more A+T-rich regions (with the peak at 44–46% G+C content for the oldest group). By contrast, young *Alu* subfamilies are found in more A+T-rich regions while the enrichment shifts to more G+C-rich regions over time.[1] Thus, although the overall distributions of *Alu* and SVA elements are both in G+C-rich region, the shifting of their G+C distribution patterns are opposite.

Several scenarios could explain the data noted above. For instance, if SVA elements prefer to insert in A+T-rich regions, the distribution of the elements may be due to positive selection acting on the young elements in the G+C-rich regions. The positive selection hypothesis requires selection to be acting on the majority of the SVA elements and will result in the fixation of the elements that are under selection in a short period of time. However, the ~30% polymorphism observed for the two human-specific subfamilies (<4 Myrs old) is consistent with neutral expectations. Therefore, positive selection is unlikely to be the causative factor for the observed scenario. Another possibility is that elements were preferentially removed from the A+T-rich regions. The removal of *Alu* elements from A+T-rich regions due to unequal homologous recombination is thought to contribute to the shifting of the *Alu* distribution pattern over time.[34,35] But given that there is one SVA every 1.03 Mb on average (compared to one *Alu* every 3 kb), the probability of a large-scale removal of the elements *via* this type of recombination seems remote. Nevertheless, due to the relatively low copy number of SVA elements, less recombination-related problems may be generated as compared to L1 elements. This may give the SVA elements a better chance to be fixed in G+C-rich regions. The possibility remains that this fixation advantage of SVA elements in G+C-rich region combining with other unknown mechanisms shifted their distribution over time.

Another possibility is that SVA elements are preferentially inserted in the G+C-rich regions. Under this hypothesis, more SVA elements would have to be removed from G+C-rich regions over time to generate the observed shift in the distribution. However, this seems counter-intuitive for at least two reasons. Firstly, SVA elements are quite possibly using the retrotransposition machinery as L1 and *Alu* elements,[2] and both young *Alu* and L1 elements are found more common in the A+T-rich regions of the genome.[1] Secondly, the G+C-rich regions are known to have higher gene densities; therefore, the insertion of retroelements in these regions will have a larger chance of influencing gene expression and causing genetic defects. Nevertheless, we have no evidence to rule out the possibility of the changing insertion preference at this moment. In addition to these scenarios, other mechanisms such as compositional matching[36] may also have played a role in shaping the SVA distribution.

Our observations in this study bring up an interesting question: if *Alu*, L1 and SVA elements are indeed all retrotransposed using the L1 enzymatic machinery, why are there such dramatic differences among their distributions? Multiple scenarios have been proposed for the shifting of the *Alu* distribution to G+C-rich regions, yet the topic remains intensively debated.[1,25,34–40] The observed distribution of the SVA elements undoubtedly added another interesting dimension to the overall retroelement genomic distribution puzzle.

The impact of SVA elements

Similar to other mobile elements, the insertion of SVA elements in the genome may have a profound impact on the genomic architecture and stability.[11,12,41] One of the properties of the SVA elements is their high G+C content. A typical full-length element has about 60% G+C content while the G+C content of the VNTR region may even exceed 70%. This makes each SVA element a potential mobile CpG island and the insertion of the element may influence the surrounding genomic environment. SVA elements are also enriched in potential functional units, including SP1 binding sites (GGCGG) in the VNTR region and hormone responsive elements (HRE) in the S part, to name a few. Since many of these units are harbored in the repetitive regions (both of the hexamer repeat region and the VNTR region), multiple copies can

be found in a single element. If inserted near a gene, SVA elements may influence the gene expression pattern. An examination of SVA insertions adjacent to the 5′ end of annotated genes indicated that approximately 200 elements integrated within 5000 bp upstream of annotated genes in the UCSC May 2004 (hg17) human genome assembly. Each of these elements has the potential to alter the transcription of the nearby genes. However, detailed studies of gene expression will be required to accurately determine the impact of these elements on gene expression.

What is certain, however, is that SVA can have a negative impact on the genome. To date, four cases of different diseases have been reported related to the SVA insertions.[2,7–9] The existence of diseases caused by SVA insertions, along with the presence of partial SVA transcripts in dbEST (data not shown), provide strong evidence for the ongoing amplification of SVA elements in the human genome. However, as a result of their relatively lower copy number (<3000), SVA elements are likely to have a much lower mutagenic recombination-based impact on the genome as compared to the more abundant *Alu* and L1 elements. In fact, only one case of disease caused by SVA-mediated recombination has previously been reported.[10]

Similar to L1 elements, SVA elements sometimes bypass their own polyadenylation site and use a downstream site, a phenomenon known as "transduction".[12,42] In our study, we found about 10% of the SVA elements also transduce a sequence at the 3′ end of the element. The transduced sequence ranged from several base-pairs to more than several thousand base-pairs, and some of them contained coding sequences. The ability of SVA elements to utilize 3′ transduction to shuffle genetic material represents yet another impact that SVA elements have on the genome. In some cases, this so-called "exon shuffling" may generate new proteins with novel functions.[42]

Evolution of SVA elements

The evolutionary history of SVA retroposons was examined using human SVA elements. Since large amounts of genomic sequence data are not yet available for all primate species, we employed a QPCR-based approach to estimate the copy number of SVA elements in other primate genomes. QPCR has been proven useful in estimating DNA copy number,[43–45] because there is a quantitative relationship between the amount of DNA target sequence and the amount of PCR product generated at any given PCR cycle prior to saturation. The copy number estimates in different primates and the age estimates of SVA subfamilies generated congruent scenarios for the evolution of the SVA elements (Figure 10): the SVA elements originated or "composed" before the divergence of all hominid primates but after the divergence of hominid and Old World primates. Following the origination, the evolutionary history of SVA elements is characterized by two major expansion periods: subfamilies SVA_A and SVA_B were expanded before the divergence of great apes (orangutan, gorilla, chimpanzee and human). Another major expansion of the SVA family occurred after the divergence of

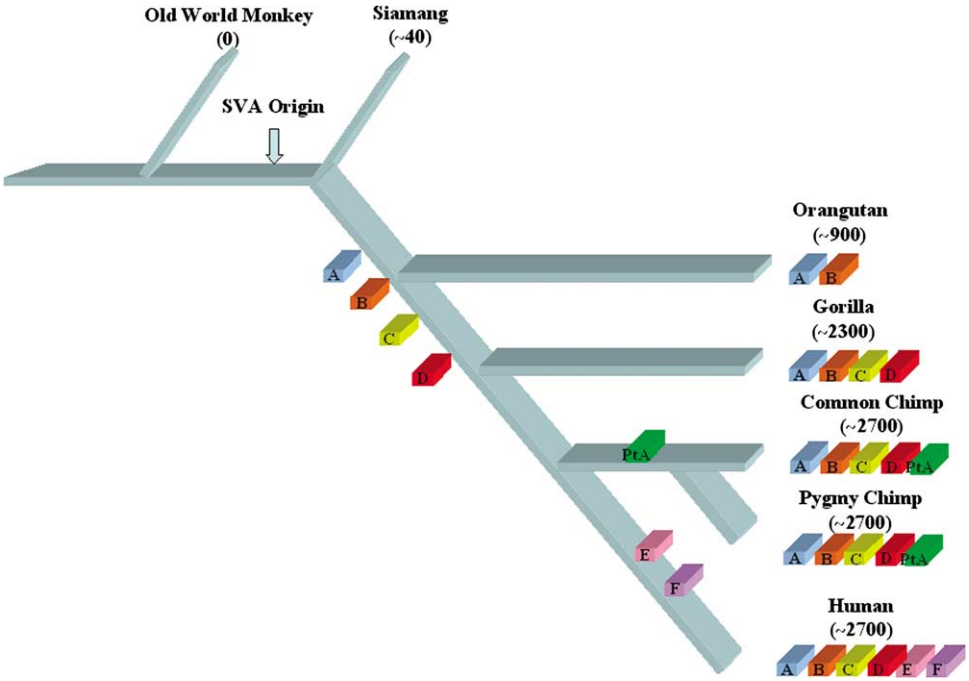

Figure 10. Amplification dynamics of the SVA family of retroposons in primates. Putative evolutionary history of the SVA lineage in primate genomes. A schematic of the hominid primate radiation is shown. The estimated copy number of SVA elements in various primate genomes is shown next to their names. The estimated time of origin and period of expansion for each subfamily are shown.

orangutan and the rest of the great apes. This major expansion is characterized by the amplification of subfamily SVA_D, which caused the number of SVA elements in human/chimp/gorilla to be much higher when compared to other species. The SVA_C subfamily may have also expanded during this period. After the divergence of gorilla from human and chimpanzee, the SVA family showed lineage-specific activity, sprouting two human-specific subfamilies. The independent expansion of multiple SVA subfamilies in the human genome indicates the existence of multiple SVA source genes, similar to *Alu* elements in the human genome.[25]

As mentioned earlier, the evolutionary history of SVA elements was examined using the human lineage as a starting point. Even though the QPCR primers were designed in a well-conserved region of the human SVA consensus, there is still insufficient data about the actual structure of SVA elements in other non-human primate genomes. The possibility remains that species-specific SVA insertions or even subfamilies exist in other hominid primates and cannot be detected using QPCR-based assays; thus, we cannot rule out that our QPCR results may underestimate the SVA copy numbers in more divergent species. Additional studies are needed to focus on the SVA elements in non-human primate genomes to augment the work presented here.

Conclusion

The SVA family of retroposons represents the third known category of mobile elements whose *de novo* mobilization results in human genetic disorders. Until now, little was known about the distribution and properties of these elements in the human genome as compared to *Alu* and L1 elements. In this study, we identified and characterized all of the SVA elements from the draft sequence of the human genome. In addition, the human genomic diversity associated with polymorphic SVA elements as well as the phylogenetic distribution of SVA elements was examined. By adding the contribution of the SVA family to our current knowledge of mobile elements, this study provides a more comprehensive picture and will further enhance our understanding of the impact of mobile elements on primate genomes.

Materials and Methods

Genome analysis

The RepeatMasker annotations of the human (hg17; May 2004 freeze) and chimpanzee genomes (panTro1; Nov. 2003 freeze) were obtained from the UCSC Genome Bioinformatics Site†. The locations of SVA elements were then extracted and inspected manually. Due to the highly variable VNTR region in SVA elements, some of the annotations did not recognize the composition of the elements correctly. In those cases, the annotations were edited manually.

Next, the SVA elements, along with 2000 bp flanking regions on both sides, were extracted from the human genomic sequence using a Perl script. The LTR-derived region of each element was then extracted manually and aligned using CLUSTAL_X.[46] For the age estimates, elements in each subfamily were aligned and subjected to further manual adjustment by removing insertions and poly(A) tails. For enhancing the quality of our data, any element that contained a deletion larger than 50 bp or could not be confidently aligned was also excluded from the alignment.

To study the SVA chromosomal distribution, human chromosome sizes and gap sizes were obtained from summary tables from the UCSC website. The information about genes, G+C content and nucleotide sequence of each chromosome was also downloaded from the same resource. For subfamily G+C analysis, the fraction of elements in each G+C bin were divided by the fraction of genome in that bin and the resulting ratio was used as measure of SVA density.[1,21] Custom-made Perl scripts were used to calculate SVA density in G+C bins as well as in genes and intergenic regions, extract flanking sequences of various sizes, extract VNTR regions from full-length SVA elements and calculate their sizes. Repeat identification in the SVA flanking regions was annotated using a local installation of the RepeatMasker program.

Statistical analysis

STATISTICA (version 6.1) was used in the statistical analysis. The chi-squared test was used to analyze chromosomal distribution of SVA elements and to compare their inter- and intragenic densities. Correlation analysis was performed to examine the relationship between SVA elements and genes/G+C content at the chromosomal level, as well as *Alu* and L1 densities in the SVA flanking regions with the G+C content. Repeat densities in the flanking regions were compared with each other using a pair-wise Student's *t*-test.

Oligonucleotide primer design and PCR analysis

Because the PCR amplicon of a typical SVA locus is usually larger than 2 kb, two separate PCRs were performed in an assay designed for L1 elements as described.[23,47] For the filled site PCR, an SVA-specific internal primer (located in the element) and a 3′-flanking unique primer were used to genotype the presence of the filled alleles of SVA insertions. For the empty site PCR, two flanking unique primers were used to genotype the empty alleles. The SVA presence/absence polymorphism can be determined by combining these two results. In this assay, 20 individuals from each of four geographically diverse human populations (European, African American, Asian and South American) were surveyed for the presence and absence of SVA elements. DNA samples from each of these populations were available from previous studies or were purchased from the Coriell Institute for Medical Research (Camden, New Jersey). The primers and annealing temperatures for each locus are shown in the supplemental Table 2‡.

† http://genome.ucsc.edu/

‡ Available at http://batzerlab.lsu.edu, under the publications section.

Other DNA samples used in this study including human genomic DNA (HeLa cell line ATCC CCL-2) and the following non-human primate species: DNA samples of *P. troglodytes* (common chimpanzee), *P. paniscus* (bonobo or pygmy chimpanzee), *G. gorilla* (western lowland gorilla), *P. pygmaeus* (orangutan) and *M. mulatta* (rhesus monkey) are available as a primate phylogenetic panel PRP00001 from Coriell. DNA samples of *H. syndactylus* (siamang) and *H. lar* (white-handed gibbon) were also purchased from Coriell (PR00721 and PR00495, respectively). DNA samples of *C. aethiops* (green monkey) were isolated from cell lines ATCC CCL70.

Quantitative PCR

SYBR® Green PCR core reagents kits were purchased from Applied Biosystems (Foster City, CA). Quantitative PCR was carried out in 50 µl of total volume reactions containing 1.25 units of AmpliTaq Gold™ DNA polymerase, 5 µl of 10× SYBR® Green PCR buffer, 3 mM $MgCl_2$, 1 mM dNTP, 0.5 mM forward (SVA_SF 5′-ACAAACACTGCGGAAGGCC−3′) and reverse (SVA_SR 5′-AGGTCTCTGGTTTTCCTAGGCA-3′) primers and 1.0 µl of DNA sample. Four serial dilutions of template DNA (10 ng, 1 ng, 100 pg and 10 pg) were used for each primate species. Amplification reactions were performed in an ABI Prism 7000 Real Time PCR System (Applied Biosystems) following the manufacturer's instructions with conditions set as follows: 95 °C, 12 min; 40 cycles of 95 °C, 15 s; and 68 °C, 1 min. A standard curve was constructed by using serial dilutions of known human genomic DNA samples ranging from 10 ng to 10 pg. The human DNA samples were included as standard with each batch of quantification reactions. The data from three identical reactions were exported from the ABI Prism 7000 System SDS Software (Applied Biosystems) into a Microsoft Excel spreadsheet and the copy number of the elements in each species was calculated based on the standard curve in each reaction. The mean values and standard deviations were calculated based on the three replications.

Acknowledgements

We thank everybody in the Batzer lab for critical reading and inspirational discussion during the preparation of the manuscript. This research was supported by the National Science Foundation grants BCS-0218338 (to M.A.B.) and EPS-0346411 (to M.A.B.); Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05 (to M.A.B.), (2000-05)-01 (to M.A.B.) and (2001-06)-02 (to M.A.B.), National Institutes of Health RO1 GM59290 (to M.A.B.) and the State of Louisiana Board of Regents Support Fund (to M.A.B.).

References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H., Jr (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* **73**, 1444–1451.
3. Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. (2004). Natural genetic variation caused by transposable elements in humans. *Genetics*, **168**, 933–951.
4. Ono, M., Kawakami, M. & Takezawa, T. (1987). A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucl. Acids Res.* **15**, 8725–8737.
5. Shen, L., Wu, L. C., Sanlioglu, S., Chen, R., Mendoza, A. R., Dangel, A. W. *et al*. (1994). Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J. Biol. Chem.* **269**, 8466–8476.
6. Callinan, P. A. & Batzer, M. A. (2005). Transposable elements and human disease. *Genome Dynam*. In the press.
7. Wilund, K. R., Yi, M., Campagna, F., Arca, M., Zuliani, G., Fellin, R. *et al*. (2002). Molecular mechanisms of autosomal recessive hypercholesterolemia. *Hum. Mol. Genet.* **11**, 3019–3030.
8. Kobayashi, K., Nakahori, Y., Miyake, M., Matsumura, K., Kondo-Iida, E., Nomura, Y. *et al*. (1998). An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature*, **394**, 388–392.
9. Rohrer, J., Minegishi, Y., Richter, D., Eguiguren, J. & Conley, M. E. (1999). Unusual mutations in Btk: an insertion, a duplication, an inversion, and four large deletions. *Clin. Immunol.* **90**, 28–37.
10. Legoix, P., Sarkissian, H. D., Cazes, L., Giraud, S., Sor, F., Rouleau, G. A. *et al*. (2000). Molecular characterization of germline NF2 gene rearrangements. *Genomics*, **65**, 62–66.
11. Batzer, M. A. & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Rev. Genet.* **3**, 370–379.
12. Ostertag, E. M. & Kazazian, H. H., Jr (2001). Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**, 501–538.
13. Grover, D., Mukerji, M., Bhatnagar, P., Kannan, K. & Brahmachari, S. K. (2004). Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. *Bioinformatics*, **20**, 813–817.
14. Carter, A. B., Salem, A. H., Hedges, D. J., Keegan, C. N., Kimball, B., Walker, J. A. *et al*. (2004). Genome-wide analysis of the human Alu Yb-lineage. *Hum. Genomics*, **1**, 167–178.
15. Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Series*, 95–98.
16. Xing, J. C., Hedges, D. J., Han, K., Wang, H., Cordaux, R. & Batzer, M. A. (2004). Alu element mutation spectra: molecular clocks and the effect of DNA methylation. *J. Mol. Biol.* **344**, 675–682.
17. Jurka, J., Krnjajic, M., Kapitonov, V. V., Stenger, J. E. & Kokhanyy, O. (2002). Active Alu elements are passed primarily through paternal germlines. *Theor. Popul. Biol.* **61**, 519–530.
18. Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

19. Goodman, M., Porter, C. A., Czelusniak, J., Page, S. L., Schneider, H., Shoshani, J. *et al*. (1998). Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* **9**, 585–598.
20. Glazko, G. V. & Nei, M. (2003). Estimation of divergence times for major lineages of primate species. *Mol. Biol Evol.* **20**, 424–434.
21. Medstrand, P., van de Lagemaat, L. N. & Mager, D. L. (2002). Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res.* **12**, 1483–1495.
22. Otieno, A. C., Carter, A. B., Hedges, D. J., Walker, J. A., Ray, D. A., Garber, R. K. *et al*. (2004). Analysis of the human Alu Ya-lineage. *J. Mol. Biol.* **342**, 109–118.
23. Myers, J. S., Vincent, B. J., Udall, H., Watkins, W. S., Morrish, T. A., Kilroy, G. E. *et al*. (2002). A comprehensive analysis of recently integrated human Ta L1 elements. *Am. J. Hum. Genet.* **71**, 312–326.
24. Salem, A. H., Myers, J. S., Otieno, A. C., Watkins, W. S., Jorde, L. B. & Batzer, M. A. (2003). LINE-1 preTa elements in the human genome. *J. Mol. Biol.* **326**, 1127–1146.
25. Cordaux, R., Hedges, D. J. & Batzer, M. A. (2004). Retrotransposition of Alu elements: how many sources? *Trends Genet.* **20**, 464–467.
26. Bandelt, H. J., Forster, P. & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48.
27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
28. Han, K., Xing, J. C., Wang, H., Hedges, D. J., Garber, R. K., Cordaux, R. *et al*. (2005). Under the genomic radar: the stealth model of Alu amplification. *Genome Res.* **15**, 655–664.
29. Hedges, D. J., Callinan, P. A., Cordaux, R., Xing, J. C., Barnes, E. & Batzer, M. A. (2004). Differential Alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14**, 1068–1075.
30. Pavlicek, A., Paces, J., Elleder, D. & Hejnar, J. (2002). Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. *Genome Res.* **12**, 391–399.
31. Lavie, L., Maldener, E., Brouha, B., Meese, E. U. & Mayer, J. (2004). The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res.* **14**, 2253–2260.
32. Volloch, V. Z., Schweitzer, B. & Rits, S. (1995). Transcription of the 5′-terminal cap nucleotide by RNA-dependent DNA polymerase: possible involvement in retroviral reverse transcription. *DNA Cell Biol.* **14**, 991–996.
33. Hirzmann, J., Luo, D., Hahnen, J. & Hobom, G. (1993). Determination of messenger RNA 5′-ends by reverse transcription of the cap structure. *Nucl. Acids Res.* **21**, 3597–3598.
34. Hackenberg, M., Bernaola-Galvan, P., Carpena, P. & Oliver, J. L. (2005). The biased distribution of Alus in human isochores might be driven by recombination. *J. Mol. Evol.* **60**, 365–377.
35. Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J. V. & Bernardi, G. (2001). Similar integration but different stability of Alus and LINEs in the human genome. *Gene*, **276**, 39–45.
36. Gu, Z., Wang, H., Nekrutenko, A. & Li, W. H. (2000). Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene*, **259**, 81–88.
37. Rynditch, A. V., Zoubak, S., Tsyba, L., Tryapitsina-Guley, N. & Bernardi, G. (1998). The regional integration of retroviral sequences into the mosaic genomes of mammals. *Gene*, **222**, 1–16.
38. Ovchinnikov, I., Troxel, A. B. & Swergold, G. D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res.* **11**, 2050–2058.
39. Brookfield, J. F. (2001). Selection on Alu sequences? *Curr. Biol.* **11**, R900–R901.
40. Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V. V. & Jurka, M. V. (2004). Duplication, coclustering, and selection of human Alu retrotransposons. *Proc. Natl Acad.Sci. USA*, **101**, 1268–1272.
41. Deininger, P. L., Moran, J. V., Batzer, M. A. & Kazazian, H. H., Jr (2003). Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* **13**, 651–658.
42. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H., Jr (1999). Exon shuffling by L1 retrotransposition. *Science*, **283**, 1530–1534.
43. Ginzinger, D. G., Godfrey, T. E., Nigro, J., Moore, D. H., II, Suzuki, S., Pallavicini, M. G. *et al*. (2000). Measurement of DNA copy number at microsatellite loci using quantitative PCR analysis. *Cancer Res.* **60**, 5405–5409.
44. Alonso, A., Martin, P., Albarran, C., Garcia, P., Garcia, O. & de Simon, L. F. *et al*. (2004). Real-time PCR designs to estimate nuclear and mitochondrial DNA copy number in forensic and ancient DNA studies. *Forensic Sci. Int.* **139**, 141–149.
45. Walker, J. A., Kilroy, G. E., Xing, J., Shewale, J., Sinha, S. K. & Batzer, M. A. (2003). Human DNA quantitation using Alu element-based polymerase chain reaction. *Anal. Biochem.* **315**, 122–128.
46. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* **25**, 4876–4882.
47. Sheen, F. M., Sherry, S. T., Risch, G. M., Robichaux, M., Nasidze, I., Stoneking, M. *et al*. (2000). Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* **10**, 1496–1508.

Edited by J. Karn

(Received 22 August 2005; received in revised form 22 September 2005; accepted 27 September 2005)
Available online 19 October 2005